Adv Data Anal Classif (2015) 9:341–365
DOI 10.1007/s11634-014-0188-0

REGULAR ARTICLE

A diffusion model for churn prediction based on sociometric theory

Uroš Droftina · Mitja Štular · Andrej Košir

Received: 4 September 2013 / Revised: 1 September 2014 / Accepted: 13 October 2014 /
Published online: 22 October 2014

Abstract Churn prediction has received much attention in the last decade. With the evolution of social networks and social network analysis tools in recent years, the consideration of social ties in churn prediction has proven promising. One possibility is to use energy diffusion models to model the spread of influence through a social network. This paper proposes a novel churn prediction diffusion model based on sociometric clique and social status theory. It describes the concept of energy in the diffusion model as an opinion of users, which is transformed to user influence using the derived social status function. Furthermore, a novel diffusion model prediction scheme applicable to a single user or a small subset of users is described: the Targeted User Subset Churn Prediction Scheme. The scheme allows fast churn prediction using limited computing resources. The diffusion model is evaluated on a real dataset of users obtained from the largest Slovenian mobile service provider, using the F-measure and lift curve. The empirical results show a significant improvement in prediction accuracy of the proposed method compared with the basic spreading activation technique (SPA) diffusion model. More specifically, our approach outperforms a basic SPA diffusion model by 116 % in terms of lift in the fifth percentile.

U. Droftina (✉) · M. Štular
Telekom Slovenije, d.d., Cigaletova 15, 1000 Ljubljana, Slovenia
e-mail: uros.droftina@telekom.si

M. Štular
e-mail: mitja.stular@telekom.si

A. Košir
Faculty of Electrical Engineering, University of Ljubljana,
Tržaška 25, 1000 Ljubljana, Slovenia
e-mail: andrej.kosir@ldos.fe.uni-lj.si

Keywords Diffusion model · Churn prediction · Sociometric clique · Social status · Telecommunications

Mathematics Subject Classification 91D30 Social networks · 62-07 Data analysis · 62P30 Applications in engineering and industry · 05C85 Graph algorithms

1 Introduction

The telecommunication sector is one of the most profitable in the world. It has been estimated that around 4.7 trillion dollars was spent in 2012 in this sector (Plunkett 2012). Therefore, it is understandable that competition within the sector is intense. Most countries have several service providers, and their common goal is to maximize possible revenue. Service providers therefore strive to acquire more and more new subscribers. In the last decade, mobile telecommunication markets became saturated in many countries; i.e., the number of applied mobile numbers reached or even exceeded the number of residents. To retain or even increase their market share, service providers are faced with two options: to retain users and to acquire users from competitive providers. Since the cost of retention is about one-sixth of that of acquisition (Rosenberg and Czepiel 1984), providers mainly focus on the first action.

Service providers in every healthy telecommunication market are faced with the possibility of their customers leaving to competitive providers on a daily basis. Users who leave one provider and transfer their business to the competition are called churners, and the action of this transfer is called churn. In many markets all over the world, the churn rate has increased significantly since the introduction of mobile number portability. The reason users transfer their subscriber numbers to other service providers is usually their dissatisfaction with their current provider or a better offer from the competition. The monthly churn rates can reach and even exceed 15 %, depending on the considered service provider and the market this service provider operates in. Although churn is an important revenue factor in the telecommunication business, other businesses, such as food-based retailing (Miguéis et al. 2012), the credit-card business (Naveen et al. 2010), online gaming (Kawale et al. 2009), and advertising (Yoon et al. 2010), also face this problem.

Telecommunication service providers possess large databases in which every user event or interaction is stored. These data are an invaluable source of discovering new knowledge about users and are also key in customer churn prediction. Two different approaches to churn prediction exist in the literature. A more widely used approach uses machine learning and data mining techniques, where different user features are used as inputs; e.g., demographic information, usage history, and payment discipline. An extensive overview of state-of-the-art classification techniques was provided by Verbeke et al. (2012). A second, more recent approach predicts churn by considering telecommunication networks as social networks, and uses social network analysis tools to find specific patterns and features in networks. A pioneering work in this area was that of Dasgupta et al. (2008), who used a spreading activation technique (SPA) diffusion algorithm to model the spread of churn influence and thus predict potential churners. This approach is based on the assumption that the users' probability to churn

will increase with the increasing number of strongly connected recent churners in their neighbourhood.

Although the SPA algorithm proves to be an effective approach with which to determine potential churners, detailed investigations have suggested possibilities of improving the model. One of the identified issues of the SPA diffusion model is that it assigns all known recent churners the same initial energy before starting a diffusion process, hence ignoring some important factors. The same initial energy reflects the same amount of churn influence for all recent churners. However, it has been shown that different users have very different influences in their corresponding social neighbourhoods (Kempe et al. 2005). Consequently, the initial energy in the diffusion process should differ among users.

Additionally, the SPA diffusion model considers the amount of communication (sum of calls) between two users as the only factor of influence between them. For example, consider a user u that communicates with users v_1 and v_2 for the same amount time in an observed time period. User v_1 is a close friend of u and thus has considerable influence on u. Users u and v_1 also have other friends in common, who are all connected with each other and together form a closely connected group. On the other hand, users u and v_2 are connected only in business terms (e.g., an auto mechanic and a customer) and do not have as much influence on each other as their friends do (see Fig. 1). We believe that in addition to the amount of communication, social factors should also be considered.

The first goal of this research is to introduce a method of finding closely connected groups of users using sociometric clique theory (Alba 1973; Luce and Perry 1949; Mokken 1979) that is based on the communication patterns of the users. These groups are used to determine new initial energy values, which are used later in the diffusion process.

The second goal of this research is to introduce a novel diffusion model based on social status theory that improves the churn prediction accuracy of the state-of-the-art

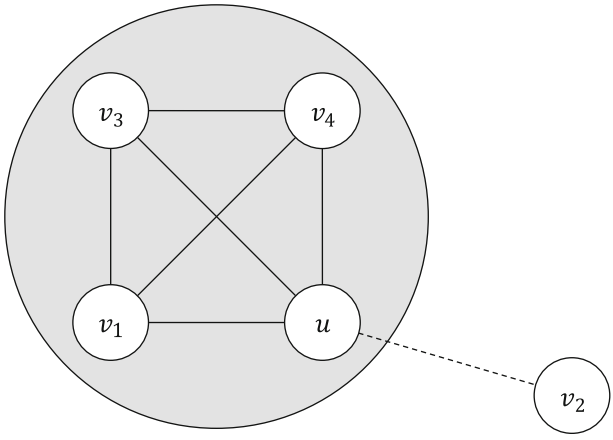

Fig. 1 Example of a closely connected user group (users u, v_1, v_3, and v_4)—a clique—and a socially less connected user v_2 outside the clique. A clique is a group of users who are all connected with each other. The clique is presented in the figure as a *grey circle*

approach. Additionally, an explanation is provided of what energy represents in the diffusion model and how it is transformed into influence. We consider energy to be equal to user opinion and influence to be a transformation of user opinion. We calculate the influence of each user as a product of the opinion and social status of the user. In the empirical part of this study, we compare the predictive power of our model with that of the SPA diffusion model (Dasgupta et al. 2008), using a proposed novel diffusion model prediction scheme that allows fast churn prediction using limited computing resources.

The reminder of the paper is structured as follows. Section 2 introduces a short review of churn prediction in the literature. Section 3 describes the proposed approach for calculating the initial energy values and a novel diffusion model algorithm. Additionally, a novel churn prediction scheme is proposed. Section 4 compares the performance of proposed diffusion models with that of the basic SPA diffusion model using real data from a Slovenian mobile service provider. Section 5 discusses the experimental results. The paper closes with a conclusion and discussion of additional open issues and plans for further work.

2 Churn prediction in the literature

There are several known approaches used in churn prediction. The first and most widely used approach employs data mining techniques and statistical analysis. For this approach, we describe some of the most referred works relevant to the proposed method. A pioneering work in this area was published by Mozer et al. (2000), where the authors explored techniques from statistical machine learning to predict churn and subsequently determine what incentives should be offered to subscribers to improve retention and maximize profitability to the carrier. The techniques include (1) logistic regression, (2) the use of decision trees, (3) the use of neural networks, (4) the boosting of decision trees with the AdaBoost algorithm, and (5) the boosting of neural networks with the AdaBoost algorithm. Experiments were based on a database of nearly 47,000 domestic subscribers in the United States and included information about their usage, billing, credit, application, and complaint history. Similar work was also done by Nath and Behara (2003), who performed customer churn analysis with the help of the Naive–Bayes algorithm, and by Ferreira et al. (2004), who evaluated different data mining methods, namely the use of neural networks, decision trees, genetic algorithms, and neuro-fuzzy systems. The whole process of predicting customer behaviour in telecommunications has been described by Yan et al. (2004). The description includes defining the predicting target, extracting customer data and errors in assembling training sets, preprocessing the raw data, and taking data characteristics into account. Since each proposed algorithm also has its drawbacks besides its advantages, no best universal single-algorithm solution has been established. Consequently, hybrid approaches that combine the advantages of different methods have been proposed; e.g., Qi et al. (2008) combined the modified *alternative decision tree* algorithm and *logistic regression*.

Data mining techniques have been widely used in churn prediction modelling. However, the problem with the data mining approach using user features is that it only considers individual attributes of users. It is believed that user decisions such

as churn are commonly affected by other connected users. This is not considered in the traditional approach employing data mining. With the evolution of social networks in recent years, considering social ties in churn prediction has proven to be a promising approach. Recent research in the area of churn prediction modelling has tried different approaches of combining feature-based techniques with social-network analysis tools. Most works make use of converting relational representation data to feature-based data and enriching "classical" features. Some works include different centralities (e.g., betweenness and closeness) as additional features, some add features such as the number of neighbouring churners and the number of calls to neighbouring churners (Dierkes et al. 2011), and others use calculated network attributes, such as the neighbour composition, tie strength, similarity, and homophily (Zhang et al. 2010, 2012). Richter et al. (2010) proposed a systematic study that evaluates the relevance of group-related features to churn.

A pioneering work in the field of examining social ties and their relevance to churn in telecommunication was published by Dasgupta et al. (2008). The authors proposed a spreading activation-based technique that predicts potential churners according to their corresponding social network, including information on users who have already churned. The idea is that users, who recently churned, influenced with their decisions other users in their social neighbourhood. The underlying algorithm assigns all recent known churners the same initial positive non-zero energy (e.g., 1) while all other users start with zero energy. An energy spreading technique is then initialized where, in each iteration, active nodes (with non-zero energy) transfer a portion of their energy to their neighbours, relative to the strength of connections between nodes. The overall amount of energy remains the same throughout the whole process. The iterating process continues until a stable state is achieved. Afterwards, a simple threshold-based technique is used where a threshold T is fixed and nodes with energy greater than T are labelled potential churners while others are labelled non-churners. Higher energy therefore presents higher propensity to churn.

This SPA algorithm is an effective approach with which to determine potential churners affected by recent churners, who are strongly connected to them. However, several detailed studies of the SPA algorithm suggested possibilities of improving the model (Baras et al. 2012; Kawale et al. 2009). Kawale et al. (2009) suggested that "churn energy" should spread through the user network in both positive and negative senses. They proposed a modified SPA diffusion model that propagates negative as well as positive influence. The model was evaluated on users of online multi-player role-playing games and found to perform better than the basic SPA algorithm. Baras et al. (2012) compared several different methods of estimating the connection strength between users, which can be used as input to an SPA diffusion model. They showed that using a social measure (i.e., the number of shared vs. unshared outgoing neighbours) as a connection weight instead of using the number of calls between users significantly improves the predictive power of an SPA diffusion model. Their experimental results were based on real mobile network data. Here, the authors manipulated the network graph of users that is used as an input to the diffusion model while leaving the diffusion model intact. Therefore, an unexplored possibility of improving the predictive power of the diffusion model lies in modifying the diffusion algorithm itself.

A diffusion algorithm has two basic parts: determination of initial values and the spreading of iterative energy. In this work, we deal with the analysis and optimisation of both parts by considering findings from sociometric theory (Moreno 1953). Although sociometric theory was established by Moreno almost 80 years ago, it is still widely explored and used among researchers nowadays. Research on social concepts such as popularity, social preference, social status, and influence is usually performed and evaluated on smaller connected groups of users; e.g., students within the same grades (Marks et al. 2012; Berg and Cillessen 2012). This research has shown that people within a closely connected group have greater influence on each other than on connected users outside their group. Moreover, they usually share the same opinion (Moody 2001). Such a closely connected group of users, where everyone knows everyone else, is called a clique (Alba 1973; Luce and Perry 1949). In this work, we use traffic data [calls and short message service (SMS) messages] of mobile-network users to create a social network graph, and by discovering cliques in this graph, we extract new information on the social structure of the network graph and use it to modify the SPA diffusion model algorithm. Although finding all cliques in a social graph is computationally expensive (with the computational time increasing exponentially with the number of connections in a graph Gary and Johnson 1979), several different algorithms have been proposed to reduce this to polynomial time (Gross and Yellen 2003). Additionally, in our work, we define the social status of users using only call logs. To the best of our knowledge, no other study has used this approach in the SPA diffusion process. Our results show significant improvement in churn prediction can be achieved using our modified diffusion model.

3 Proposed solutions to the SPA diffusion model

In this section, the basic reasoning and methodology of the novel diffusion model are presented. The section is composed of three distinct parts: determination of new initial values using sociometric clique theory (Alba 1973; Luce and Perry 1949), modification of the SPA diffusion model algorithm that considers additional properties of the social graph and the social statuses of users, and the proposal of a novel diffusion model prediction scheme, the targeted user subset churn prediction scheme (TUSCPS), which allows fast churn prediction using limited computing resources.

3.1 Basic reasoning toward the proposed model

Service providers have many different users, where each user u has an opinion about its service provider. The opinion is a complex concept, but we simplify it by encoding it into a one-dimensional real number called energy $E(u)$. As in the basic SPA diffusion model algorithm (Dasgupta et al. 2008), higher values of energy reflect a higher propensity to churn. In contrast to the SPA diffusion model, where only non-negative energy values are used, we also use negative energy, which reflects a positive opinion on a service provider, as described by Kawale et al. (2009). Values of energy around zero represent neutral opinion.

We assume energy (opinion) is spread among users through a social network according to the proposed diffusion model (see Sect. 3.3). Therefore, the user's opinion is changing because of influence from connected users in his/her social network neighbourhood. At each iteration step k of the diffusion model evolution, the energy of a user u is denoted $E(u, k)$. The energy evolves from initial values of energy $E(u, 0)$ determined for each user prior to building a diffusion model (see Sect. 3.2). Initial values of energy $E(u, 0)$ present opinion at the present time. We obtain the influence from opinion by multiplying the opinions of users with the social statuses of users (see Sect. 3.3.2). At the end of diffusion model building, each user is assigned a certain value of churn influence in the form of energy. We predict that user u will churn if his/her energy $E(u)$ is higher than a certain energy threshold T at a given step of evolution k.

Building of a model is an iterative process where, in each iteration, influence is spread between each pair of users in a network. However, our preliminary results show that best results are not achieved in a steady state, like the results in Dasgupta et al. (2008), but rather after just a few iterations of model building. The number of iterations can be interpreted as the number of degrees away from the source that an influence can reach (rather than the number of time steps). However, we believe that the average user can influence only the first few orders of users, with the greatest influence being on their direct neighbours (see Sect. 3.3.4). We determine the optimal number of iterative steps for different diffusion models in the empirical part of this work.

The goal of the proposed method is to predict churners for a future time interval. First, churner data are extracted from the training time interval I_t and used to build a diffusion model. Then, the model is evaluated on a subsequent time interval I_e, referred to as the evaluation interval. We assume basic opinion dynamics is preserved from the beginning of I_t to the end of I_e. Clearly, this assumption is true only when the total time span $|I_t \cup I_e|$ is sufficiently short. However, the training time interval must also be sufficiently long to gather enough user data. Our experimental results show that applicable values of time interval spans are $|I_t| = 2$ months and $|I_e| = 1$ month. These values are used in the empirical part of this research.

3.2 Model for determining initial values using sociometric clique theory

Social science theory states that people within a closely connected group have greater influence on each other than on connected users outside their group (Berg and Cillessen 2012). Moreover, they usually share the same opinion (Moody 2001) and thus should have similar initial energy. Several different definitions of a clique exist in the area of graph theory. We follow the definition established by Luce and Perry (1949), where a clique is a group of at least three nodes that form a maximal complete subgraph. One of the goals in this work is to propose a realistic approach of determining the initial energy of users, according to their direct connections, with a focus on cliques. We use cliques to model groups of users who all know each other.

Below, we present our reasoning toward the determination of initial energy values for each user. We also provide the whole initial energies determination procedure in a more concise form in Algorithm 1.

Algorithm 1: Initial energies determination pseudo-code

Data: $G = (U, E)$, where G is an undirected graph, comprised of vertices (users) U and edges (connections) E.
Result: Initial energy values $E(u, 0), u \in U$.
begin
  find a set of all maximal cliques Q in graph G;
  foreach $u \in U$ **do**
    set $E_s(u,0) = \begin{cases} 0, & u \in C; \\ 1, & u \notin C. \end{cases}$ `// self contribution`
    if $u \in Q$ **then** `// if u a member of any clique Q`
      foreach $q \in Q$ *where* $u \in q$ **do**
        set $E_{c,i}(u,0) = 2\frac{n_{cq,i}(u)}{n_{q,i}(u)} - 1$; `// contribution for each clique`
      end
      set $E_c(u,0) = \frac{1}{m_u}\sum_{i=1}^{m_u} E_{c,i}(u,0)$; `// clique contribution`
      set $E_o(u,0) = 0$; `// out-of-clique contribution`
    else
      set $E_c(u,0) = 0$; `// clique contribution`
      set $E_o(u,0) = 2\frac{n_{co}(u)}{n_o(u)} - 1$; `// out-of-clique contribution`
    end
    set $E(u,0) = E_s(u,0) + E_c(u,0) + E_o(u,0)$; `// initial energy`
  end
end

3.2.1 Proposed model for determining initial values

Consider U to be the set of all observed users in a dataset. Some of the users in U are churners. We denote a set of recent churners as C, where $C \subseteq U$. We define three types of contributions to initial energy:

1. The self contribution,
2. The clique contribution, and
3. The out-of-clique contribution.

The self contribution is the same energy contribution as in the basic SPA model; i.e., all recent churners are assigned the same non-negative self-contribution energy $E_s(u, 0)$, while all other users are assigned $E_s(u, 0) = 0$. We define a unit of energy by assuming that the energy of a churner $u \in C$ is $E_s(u, 0) = 1$. We present this as follows.

$$E_s(u,0) = \begin{cases} 0, & u \in C; \\ 1, & u \notin C. \end{cases} \tag{1}$$

The clique contribution is an energy contribution of all cliques a user is a part of. We denote these cliques by $q_1(u), q_2(u), \ldots, q_{m_u}(u)$, where m_u is the number of all cliques a user u is a member of. The clique contribution for each clique is calculated using Eq. (2), where $n_{cq,i}(u) = |q_i(u) \cap C|$ is the number of all churners in a clique $q_i(u)$, and $n_{q,i}(u) = |q_i(u)|$ is the number of all users in a clique $q_i(u)$. $E_{c,i}(u, 0)$ can take any value in the interval $[-1, 1]$ where the value -1 describes a clique without any churners, and the value 1 describes a clique containing only churners. If a user

is a member of more than one clique then the final clique contribution is calculated as the average of the individual clique contributions, as described by Eq. (3). Clique contribution equations are derived from basic assumptions of social structure theory (Blau 1975; Berg and Cillessen 2012).

$$E_{c,i}(u,0) = 2\frac{n_{cq,i}(u)}{n_{q,i}(u)} - 1 \tag{2}$$

$$E_c(u,0) = \frac{1}{m_u}\sum_{i=1}^{m_u} E_{c,i}(u,0) \tag{3}$$

The out-of-clique contribution is the energy contribution of users who are not members of any clique. It is calculated using Eq. (4), where n_{co} is the number of all churner neighbours, and n_o is the number of all neighbours of a considered user. Clique and out-of-clique contributions are mutually exclusive.

$$E_o(u,0) = 2\frac{n_{co}(u)}{n_o(u)} - 1 \tag{4}$$

Finally, initial energy values are calculated as a sum of all three contributions (5). The sign of initial energy symbolically represents positive or negative churn opinion.

$$E(u,0) = E_s(u,0) + E_c(u,0) + E_o(u,0) \tag{5}$$

3.3 SSA–SPA diffusion model

Our algorithm for assigning initial energy values considers the user's own state (churner or non-churner), the states of his/her first-order neighbours, and all the cliques a user is a member of. Therefore, the initial values themselves can already be used as churn prediction scores. However, some influence can still come from indirectly connected neighbours; therefore, a diffusion algorithm should still be applied. In this work, we also propose a new improved version of the SPA diffusion algorithm that is based on the social structure theory (Blau 1975; Berg and Cillessen 2012) and the modified social status theory relating to sociograms (Marks et al. 2012; Moreno 1953). We adopt the theory of social status (Marks et al. 2012; Moreno 1953; Berg and Cillessen 2012) and refer to our proposed algorithm as the Social-Status-Aware SPA model (SSA–SPA).

We assume that opinion (energy) is exchanged between users proportionally to the influence users have on each other. This can be presented with Eq. (6), where $i(u \to v, k)$ is an influence function in step k, $ss(u \to v)$ is a pairwise social status, and f is a mapping function that adapts social status values as explained in Sect. 3.3.2. In simple words, influence equals the opinion difference multiplied by the mutual social status. An example of the social graph of two directly connected users u and v with their connections is presented in Fig. 2.

$$E(u \to v, k) = i(u \to v, k)\, f(ss(u \to v)) \tag{6}$$

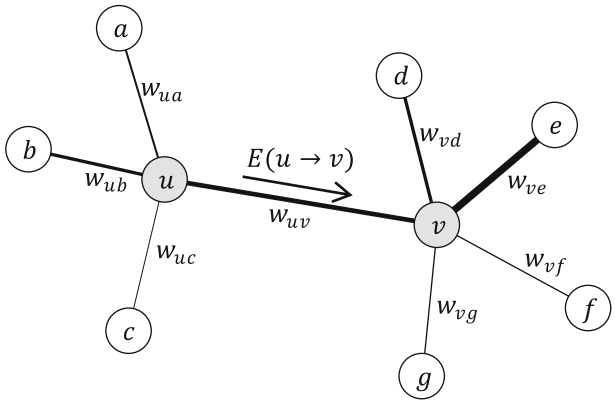

Fig. 2 An undirected graph of connected users u and v, and their connected neighbouring users. Label w_{uv} presents a weight of connection between users u and v, which is also symbolically represented by the width of the *connection line*. $E(u \rightarrow v)$ denotes energy transfer from user u to user v

3.3.1 Influence function

The first factor in the energy transmission function is the influence function, which can be expressed as Eq. (7), where d is a spreading factor, α a transfer function, and $E(u, k-1)$ and $E(v, k-1)$ respectively energy values of users u and v in iteration step $k-1$. Transfer function α is a function that normalizes energy transfers between nodes according to the node neighbourhoods in an underlying social graph. The method of normalization can affect the prediction accuracy of a built model. In the basic SPA algorithm [Eq. (8)], $\alpha(u \rightarrow v)$ is equal to a weight of a connection between users u and v, normalized to the sum of all connections of a transmitting user (seed user). Here we introduce a general normalization neighbourhood of a transmitting user $N_t(u)$ and the neighbourhood of a receiving user $N_r(v)$. These neighbourhoods can be a singleton $N_0(u) = \{u\}$, standard neighbours $N_1(u) = \{v + \text{other standard neighbours}\}$, a sub-graph of second-order users $N_2(u) = \{v + \text{other standard neighbours} + \text{second order neighbours}\}$, or a sub-graph of higher-order neighbours. A clique could also be used here (see Sect. 3.2.1).

$$i(u \rightarrow v, k) = d\alpha(u \rightarrow v)(E(u, k-1) - E(v, k-1)) \tag{7}$$

$$i(u \rightarrow v, k) = d\alpha(u \rightarrow v)E(u, k-1) \tag{8}$$

Our preliminary testing of several different variations of transfer function on a real dataset yielded the best results by setting α as the number of mobile events between users u and v relative to the sum of all mobile events of users u (energy-transmitting user) and v (energy-receiving user), as shown by Eq. (9). This transfer function is also used in the empirical part of this paper (see Sect. 4).

$$\alpha(u \rightarrow v) = \frac{w(u, v)}{\sum_{n_u} w(n_u, u) + \sum_{n_v} w(n_v, v) - w(u, v)} \tag{9}$$

The transfer function in the basic SPA diffusion model shows that energy transfer is dependent on the value of energy of the transmitting user (u) in a previous iteration step $k-1$. We believe that this is not an optimal solution. For example, if users u and v both have positive energy (negative opinion) in iteration step $k-1$ and $E(u,k-1) < E(v,k-1)$ then, according to Eq. (8), user u would even add additional positive energy to user v, despite the fact that user u has less energy than user v, and consequently less propensity to churn. We thus propose a different approach where spreading energy in step k between users is dependent on the difference in energies between users u and v in step $k-1$, and not only on the absolute value of energy of the transmitting user.

3.3.2 Social status

The second factor in the energy transmission function is a social status function, the role of which is the transformation of opinion (energy) into influence (changed opinion). There are several different possibilities of measuring and defining pairwise social status $ss(u \rightarrow v)$ from user u to v. Service providers have many different types of information on users (e.g., demographic data, usage history, and payment discipline) that could be used to measure social status. However, not all of these types of information are known for all users. In the case of prepaid users, the only data available is the detailed traffic data (e.g., calls, SMS use, and data usage). Therefore, to calculate the social status of as many users as possible, we must derive social-status equations from these data. Our definition of the social status of users is based only on calls between users.

We believe that the social status of users can only be determined by considering important calls; therefore, it is crucial to distinguish between important and unimportant calls. Prices of phone calls are usually directly proportional to call duration. We believe an average caller would be prepared to pay more for his/her call only if a call was important to him/her. Consequently, we consider longer calls important. On the other hand, the price of short calls is relatively low and can therefore be important or unimportant. Since we have no way of determining this, we only focus on long calls. Usually, long calls occur for two different reasons: (1) the caller is asking the callee for advice, and (2) the caller is calling a person important to him/her (e.g., a partner or family member) for the purpose of staying in touch. In both cases, a caller (user u) is calling a person that is important to him/her and considers his/her advice (user v). By doing so, user u is increasing the social status of user v and consequently decreasing his/her own social status. Considering this, we calculate social status using Eq. (10), where $n_c(u \rightarrow v, \text{long})$ represents the number of long calls from user u to v, and similarly, $n_c(v \rightarrow u, \text{long})$ represents the number of long calls from v to u. If a connection exists between users u and v but no long calls are made between them, we consider the social statuses of users u and v as equal.

$$ss(u \rightarrow v) = \frac{n_c(v \rightarrow u, \text{long})}{n_c(u \rightarrow v, \text{long})} \tag{10}$$

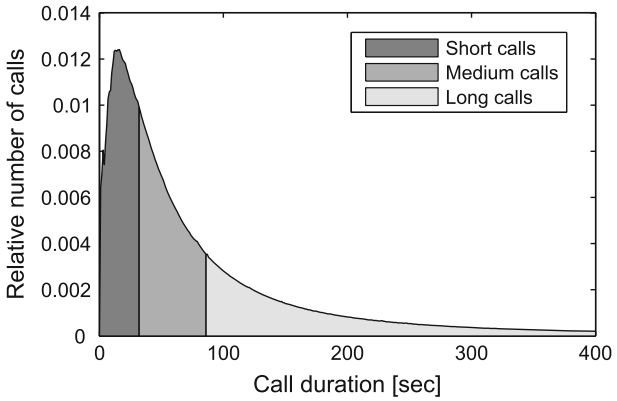

Fig. 3 An example of a histogram of call durations based on real data. All calls are classified into three groups according to call duration. The figure reveals a limit between short and medium calls at 32 s, and a limit between medium and long calls at 86 s. Only long calls are considered in calculating social status

The only parameter missing is a call duration limit between long calls and other, shorter calls. We classify all calls by duration into three classes: (1) short calls, (2) medium-length calls, and (3) long calls. Intuitively, we set an equal number of calls in each of the classes. We order all calls from the call log by their increasing duration and classify the first third of the calls as short calls, the second third of the calls as medium-length calls, and the remaining third of the calls as long calls. Duration thresholds between these three call classes are clearly seen on a histogram of call durations in Fig. 3 for a real data sample of more than 63 million calls obtained from a Slovenian mobile-service provider. The figure reveals a limit between short and medium length calls at 32 s, and a limit between medium-length and long calls at 86 s. These values are also used in the empirical part of this work.

3.3.3 Social status function

Social status ss is defined in such way that user v has higher social status than u if $ss(u \to v) > 1$ and lower social status if $ss(u \to v) < 1$. If $ss(u \to v) = 1$, then users u and v have equal social status. If $n_c(v \to u, \text{long}) \gg n_c(u \to v, \text{long})$, then $ss(u \to v) \gg 1$, which can cause transfers of energy to diverge and prevent the building of the model. We therefore need to introduce a social status function that limits the upper bound values of ss. We impose the requirements for social status function f as follows.

1. For obvious reasons, f is a continuous and increasing function; then
2. If $ss(u \to v) = 0$, then $f(ss(u \to v)) = 0$,
3. If $ss(u \to v) = 1$, then $f(ss(u \to v)) = 1$, and
4. If $ss(u \to v) = \infty$, then $f(ss(u \to v)) = 2$.

According to the above requirements, we derive social status function Eq. (11).

$$f(ss(u \to v)) = \frac{4}{1 + e^{-ss(u \to v)}} - 2 \tag{11}$$

A full equation for the energy value of user v in iteration step k is given as Eq. (12), where the new energy value of user v is a sum of energy transfers of all neighbours u_i of user v and the value of energy of user v in the previous iteration $k-1$. In the basic SPA diffusion model, each energy transfer was performed in a way that, by transferring energy e from user u to user v, the energy of user v increased by e and the energy of user u decreased by e, as a consequence of an energy (opinion) conservation law. This can be understood as an opinion change of user u, because he/she influenced user v. We see no reason for an energy conservation law and therefore do not apply one (i.e., we do not subtract the energy of the user transmitting energy).

$$E(v,k) = E(v,k-1) + \sum_{i=1}^{m} d\alpha(u_i \rightarrow v) f(ss(u_i \rightarrow v))(E(u_i,k-1) - E(v,k-1) \tag{12}$$

3.3.4 Number of iterations and energy threshold

The SPA diffusion model assumes the iterating process is carried out until a steady state is established. The number of iteration steps can be understood as the number of degrees from the source (e.g., a churner) an influence can reach. However, we believe that average users can only influence the first few orders of users, with the greatest influence being on their direct neighbours with whom they communicate directly. We support this argument by creating a bar plot of churn rate vs. the number of degrees apart from the nearest recent churner in a call network (Fig. 4). The bar plot is created using real churn data from the Slovenian mobile service provider. It shows a significantly greater churn rate for users who are directly connected to recent churners. Users who are separated from recent churners by two or more degrees display approximately the same churn rate, regardless of the degree. Therefore, the most realistic diffusion model should be achieved within first few iterations, where only the first few orders of neighbours are affected, and not when a steady state is established.

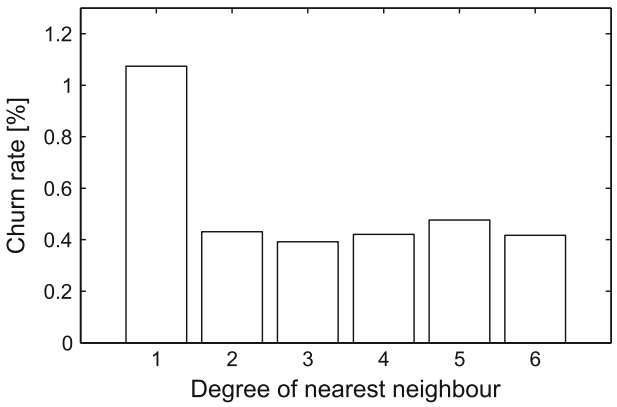

Fig. 4 Churn rate vs. degree of the nearest churning neighbour. A considerably greater churn rate is observed when a user is directly connected to at least one churner

Note that one step of the evolution of the diffusion model is one step of the underlying algorithm, and should not be confused with a real time interval such as one week. Therefore, the evolution of the diffusion model does not follow real-time transmission of opinion among connected users.

Values of energy, in any iteration, can be used as churn prediction scores. By determining an energy threshold T (which is a continuous real number), all users with energy above the threshold are predicted as churners and users with energy below the energy threshold as non-churners. An evaluation metric such as an F-measure or lift curve could be used to evaluate the result. Additionally, the threshold could also be determined as user-dependent and not equal for all users ($T = T(u)$). Using a set of user features (e.g., demographics, usage history, and connectivity features) and a continuous dependent variable regression model (e.g., linear regression), it would be possible to train a model that would determine different energy thresholds for single users or small groups of users. If user u's energy e_u is greater than energy threshold $T(u)$ after building the diffusion model, then a user would be predicted as a churner; otherwise he/she would be predicted as a non-churner. However, because of the preliminary results and for reasons of simplicity, we consider the energy threshold as being user-independent in this work.

The optimal iteration step K_{opt} and the optimal energy threshold T_{opt} can be determined using a training set. We assume that these values are time invariant and the values estimated on a training set are thus also applicable on a test (evaluation) set.

3.4 Diffusion model prediction scheme

A diffusion model is built using a graph G of connected users. However, by increasing the number of nodes and connections in G, the execution time of model building increases exponentially. Therefore, only a small subset of connected users can be evaluated at one time. Using a subset graph of users $G_s \subset G$ in building the diffusion model, there is error for users on the edge of the graph, where connections of these users with users outside the graph G_s are cut off. Therefore, the calculation of energy for users on the edge of the graph cannot be valid. An example of a graph with connections cut off on the edge of the graph is shown in Fig. 5.

Therefore, a different approach to this problem should be considered. We propose a novel diffusion model prediction scheme, applicable to an optional number of users. We first select a subset of users S for evaluation. For each user u from S, we build a network of his/her neighbourhood $N(u, d)$ up to degree d. A graph of this network is used to build a diffusion model for K iteration steps. K must be high enough so that the influence from users more than K degrees apart from user u can be ignored, while it must be small enough for the model to be built in reasonable time with available computing power. At the end of the execution of the diffusion algorithm, user u's energy is updated in a user table. When energy is determined for all users from S, a threshold method is used to predict potential churners. The threshold can be obtained from a training dataset. A great advantage of this approach is the possibility of calculating the energy for each user from S in parallel. Diffusion-model building can therefore be run on an inexpensive desktop computer instead of a supercomputer.

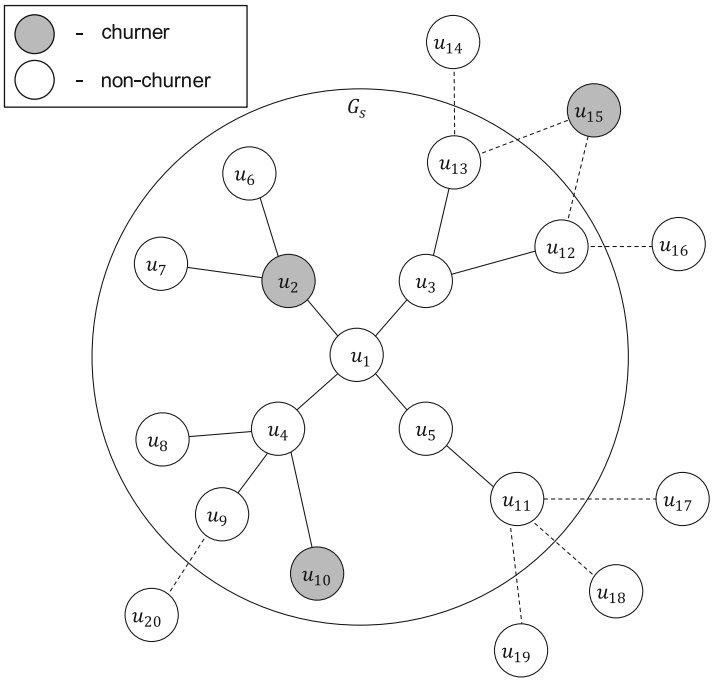

Fig. 5 An example of a subgraph $G_s \subset G$ of users (inside a *circle*) used in building a diffusion model. Because of the ignored users just outside the subgraph G_s (e.g., users u_{14}, u_{15}, and u_{16}), a churn prediction error can occur, especially for users with several connections outside G_s (e.g., users u_{11}, u_{12}, and u_{13})

Another advantage of this approach is the possibility of running diffusion-model building on any subset of users, which don't have to be connected with each other within this subset. It is reasonable to select such a targeted subset of users who are already more prone to churn; e.g., friends of recent churners or recent unsatisfied callers to customer service support. We refer to this novel prediction scheme as the TUSCPS.

4 Materials and methods

In this section, we present experimental results obtained with different prediction models, including the SPA diffusion model with proposed modifications.

4.1 Data

Experiments were conducted on real data obtained from the largest Slovenian mobile telecommunications service provider with over one million users. Since the introduction of mobile number portability in 2006, the considered provider has suffered an average annual churn rate of roughly 4 %.

To evaluate the SSA–SPA diffusion model, the considered service provider provided us with real data extracted from its data warehouse. The data included call detail records (CDRs) of natural postpaid users from July 2012, or more specifically, a log of events (calls and SMS messages) in the considered mobile network. Each event record in CDR data included the anonymised ID of party A (the caller or sender of the SMS message), anonymised ID of party B (the callee or receiver of the SMS message), time stamp of the event, and the duration, if the event was a call. These data were used to construct an undirected social graph with users of the considered service provider as nodes and communication between users as connections weighted by the number of events. The obtained social graph contained 465,000 nodes and 2.2 million connections.

Additionally, subscriber number transfer records were provided for the three months following the month of the CDR data. These data were used to label churn events. All users in the social graph that churned in August and September 2012 were labelled as seed churners (training interval), while churners from October 2012 were used in the evaluation of models as true churners (evaluation interval).

4.2 Evaluation procedure

As an appropriate measure for the evaluation of the diffusion model, we selected the F-measure (Powers 2011), which can be used to objectively evaluate skewed class classification problems, such as churn prediction. In phase 1 of the experiment, the goal is to determine the optimal number of iteration steps K_{opt} of the algorithm and the optimal threshold T_{opt}, where the best F-measure value was achieved. After each iteration of the diffusion process, energies of users are distributed on the interval $[E_{min}(k), E_{max}(k)]$. From this interval, the optimal threshold T_{opt} is determined such that the highest F-measure in the concerned iteration step k is achieved. As noted in Sect. 3.3.4, the threshold is determined to be the same for all users. The F-measure is calculated using data on true churners from the evaluation interval. After the diffusion model is built, all highest F-measures from each iteration step are used to find optimal iteration step K_{opt} where the highest of the highest F-measures was achieved. Such an F-measure value presents as an upper performance limit that real-case scenarios try to achieve. The determined optimal energy threshold T_{opt} and optimal iteration step K_{opt} are used in a real-case scenario in phase 2 of the experiment, where results of different diffusion models are evaluated using the F-measure and lift values. Additionally, upper performance F-measure limits are extracted for the dataset in phase 2 to determine how close our models approach the upper limit of the F-measure. In our experiments, the diffusion process was run for 30 iterations.

4.3 Experiment

Using a social graph and churn labels, we tested four diffusion models:

1. The SPA diffusion model (using basic initial values),
2. The SPA diffusion model (using clique-based initial values),
3. The SSA–SPA diffusion model (using basic initial values), and

4. The SSA–SPA diffusion model (using clique-based initial values).

The SPA diffusion model is that proposed by Dasgupta et al. (2008). The second model uses the SPA diffusion algorithm but clique-based initial values instead of basic initial values (see Sect. 3.2). Parameters of both algorithms were set identical to the settings proposed in Dasgupta et al. (2008). The third and fourth models are SSA–SPA diffusion models with modifications proposed in this paper (see Sect. 3.3). The third model uses basic initial values and the fourth model uses clique-based initial values.

All the algorithms were developed in Matlab and executed on a standard desktop computer (2.4 GHz Quad core, 4 GB RAM). Performance analysis showed a bottleneck in the clique calculation algorithm, where execution time increased exponentially with the increasing number of connections. Searching of cliques finished in reasonable time (under 5 min) if there were $<$17,000 connections in a graph.

The experiment was conducted in two phases. The first phase represented a training part where optimal iteration step K_{opt} and optimal energy threshold T_{opt} were set. These parameter values were used in the second phase of the experiment where different diffusion models were evaluated.

4.3.1 Phase 1: finding the optimal threshold and optimal number of iteration steps

In the first phase of the experiment, all four variants of the diffusion model were built on a predefined dataset to determine best iteration step K_{opt} and energy threshold T_{opt} for each diffusion model. The dataset was selected as follows.

1. All churners from evaluation interval I_e were ordered by decreasing number of churning neighbours and increasing number of all neighbours.
2. The first 10 churners in this order were determined as "top churners" and the basis of experimental graph G.
3. All neighbours of "top churners" up to the third degree were also added to G.
4. Nodes were connected with undirected edges, weighted by the number of mobile events (calls and SMS messages) between users.
5. Edges were established only if there were at least five mobile events between two users and there was at least one mobile event in each direction (i.e., user u called user v at least once and vice versa).

The final graph G contained 4,238 nodes and 5,471 edges. Of these users, 59 churned in the training interval (seed churners) and 31 users churned in the evaluation interval. These are the users whom the model tries to predict.

To execute the proposed SSA–SPA diffusion algorithm, the threshold of a minimum duration of long calls must be determined for the considered service provider, as described in Sect. 3.3.2. Using a histogram of durations of all calls obtained from call logs, the threshold between medium and long calls for the considered service provider was set at 86 s. This value was used as a parameter in the social status function. The setting of the diffusion factor was the same in SPA and SSA–SPA diffusion models; $d = 0.7$. Preliminary tests showed that values of d below 0.7 (down to a reasonable value of $d = 0.5$) produced similar results in terms of the F-measure, but required more iterations. Conversely, values of d above 0.7 produced slightly better but less stable results, i.e., evaluation results were very inconsistent regarding the best F-measure as

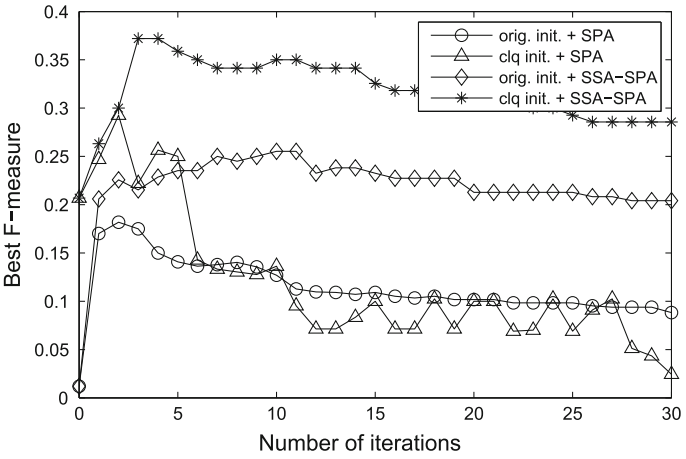

Fig. 6 Best F-measure value after each iteration for four diffusion models in phase 1 of the experiment. F-measure values on this plot are presented as the upper performance limit values that can be achieved using an optimal energy threshold in each iteration. The figure shows that the best results in all iterations are achieved using the SSA–SPA diffusion model with clique-based initial values—our proposed approach

Table 1 Optimal F-measure scores for different diffusion algorithms achieved using threshold T_{opt} after iteration step K_{opt}

	K_{opt}	T_{opt}	best F-m
orig. init. + SPA	2	0.16	0.182
clq init. + SPA	2	0.02	0.293
orig. init. + SSA–SPA	10	0.32	0.255
clq init. + SSA–SPA	3	0.04	0.372

a function of iteration steps. Therefore, as a compromise, a value of $d = 0.7$ was used, similar to the optimal value determined by Dasgupta et al. (2008).

The optimal energy threshold T_{opt} (the same for all users) and optimal iteration step K_{opt} were determined using the evaluation procedure described in Sect. 4.2. The resulting diagram of the best F-measure in each iteration step, for each of the four different diffusion models, is shown in Fig. 6. Best F-measure values for each diffusion model with their corresponding iteration step K_{opt} and energy threshold T_{opt}, used in experiment phase 2, are presented in Table 1. Note that the best F-measure values give the upper limit that the algorithm can achieve, and not what the algorithm actually achieves in a real-world scenario.

4.3.2 Sensitivity analysis of the results

A question arises as to the sensitivity of the results to changes in parameter values of: (1) the number of top churners for building a social graph, (2) the minimal number of mobile events (number of calls and SMS) for a connection, and (3) the duration threshold between medium and long calls. We performed a sensitivity analysis of the models using a one-factor-at-a-time approach (Saltelli et al. 2008) (i.e., altering one

parameter while fixing the others at their default values), and observed the values of the optimal energy threshold T_{opt}, optimal iteration step K_{opt}, and the F-measure achieved with T_{opt} and K_{opt}.

In the first step of this sensitivity analysis, we varied the number of top churners. This led to significantly different sizes of social graphs, but the values of T_{opt} and K_{opt} remained virtually constant. The order of the diffusion models in terms of their performance also remained the same. By varying the minimal number of mobile events between two users to consider them connected, we changed the density of the social graph. Consequently, the diffusion of the energy itself was also altered during the experiment. As a result, the optimal values of T_{opt} and K_{opt} varied noticeably for each of the four diffusion models, but the order of the models in terms of performance remained the same. Finally, by varying the duration threshold between medium and long calls, we indirectly changed the social status values. Despite the fact that the resulting values of T_{opt} and K_{opt} were also changing, the F-measure values deviated at most 12 % from those obtained using the default parameter set. The order of the diffusion models in terms of performance again remained constant. Overall, the sensitivity analysis showed that the number of top churners, minimal number of mobile events, and threshold between medium and long calls have an effect on T_{opt}, K_{opt}, and the F-measure, but this effect is not strong enough to change the order of the diffusion models in terms of the best F-measure.

4.3.3 Phase 2: evaluation of the diffusion model on neighbours of past churners

In the second phase of the experiment, we evaluated and compared the performance of all four variants of diffusion models. Diffusion models were built using best iteration steps K_{opt} and optimal thresholds T_{opt} obtained from the first phase of the experiment. The model was built using TUSCPS, described in Sect. 3.4. The dataset of users that a model was built on included all postpaid natural users who were neighbours of churners from August and September 2012 (the training interval). Prediction results were evaluated using churner data from October 2012 (the evaluation interval). The number of users included in the dataset was 9,958. Of these users, 107 actually churned. These are also the users whom our models tried to predict. Since four diffusion models were built for each of these 9,958 users, the execution time was a little over 70 hours. The results were evaluated using the F-measure and lift curve.

Table 2 shows the confusion matrix components (the number of true negatives, false negatives, false positives, and true positives) of the churner class for each of the diffusion models. Here, the churners are presented as positives and the non-churners as negatives. The confusion matrix values were calculated using the energy threshold T after iteration K of the diffusion process, and used to determine precision, recall, and F-measure values (Powers 2011). To determine whether the energy threshold and best iteration step are appropriate, a graph of the best F-measures after each iteration is shown in Fig. 7. Finally, the F-measure values achieved using K and T from phase 1 are compared to the best F-measure values achieved using K_{opt} and T_{opt}. Results of the comparison for each of the four diffusion models are presented in Table 3.

Table 2 Evaluation results for phase 2 of the experiment

	K	T	tn	fn	fp	tp	prec.	recall	F-m
orig. init. + SPA	2	0.16	7381	61	2470	46	0.018	0.43	0.035
clq init. + SPA	2	0.02	9541	92	310	15	0.046	0.14	0.069
orig. init. + SSA–SPA	10	0.32	9098	87	753	20	0.026	0.19	0.045
clq init. + SSA–SPA	3	0.04	9436	86	415	21	0.048	0.20	0.077

Each of the four diffusion models was built using threshold T after iteration K of the diffusion process, on a dataset of 9,958 neighbours of past churners, and then evaluated using the confusion matrix components (true negatives, false negatives, false positives, and true positives), precision, recall, and F-measure

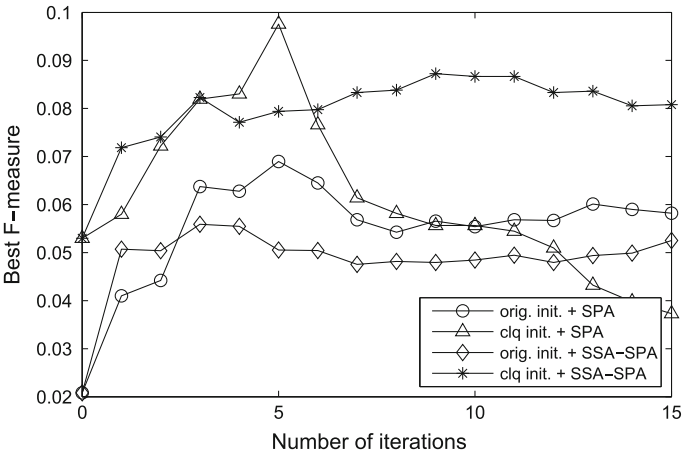

Fig. 7 A plot of upper performance limit values of the F-measure after each iteration for four diffusion models in phase 2 of the experiment. The plot reveals that the best F-measure value here could have been achieved using basic SPA with clique-based initial values if the optimal iteration step and optimal energy threshold were used. However, in a real-world scenario, best results are achieved using SSA–SPA and clique-based initial values, as is seen in Table 3

Table 3 Comparison of the F-measure achieved with the predetermined threshold T after iteration step K, and the best F-measure achieved with the optimal threshold T_{opt} after iteration step K_{opt}

	Used parameters			Best parameters			$\frac{\text{Used F-m}}{\text{Best F-m}}$ (%)
	K	T	F-m	K_{opt}	T_{opt}	F-m	
orig. init. + SPA	2	0.16	0.035	5	0.37	0.069	51
clq init. + SPA	2	0.02	0.069	5	0.01	0.098	70
orig. init. + SSA–SPA	10	0.32	0.045	3	0.46	0.056	80
clq init. + SSA–SPA	3	0.04	0.077	9	0.13	0.087	89

Results show that the SSA–SPA diffusion model using clique-based initial values came closest to the optimal prediction result

5 Discussion

Phase 1 of the experiment was conducted to retrieve an optimal combination of the energy threshold T_{opt} and iteration step K_{opt}, for which the best F-measure value is achieved. Three of the four models achieved their best F-measure value within three iteration steps, which confirms the hypothesis that churn influence does not travel more than the first few orders of degree.

Energy thresholds and iteration steps were used in phase 2 of the experiment where SPA diffusion models were evaluated. Comparison of F-measure values in Table 2 shows that all three diffusion models with included modifications outperformed the basic SPA diffusion model. The best result was achieved with the SSA–SPA model using clique-based initial values, which outperformed the basic SPA diffusion model by 120 %, if comparing F-measure values $\left(\frac{0.077-0.035}{0.035}\right)$. To see if the energy threshold and best iteration step selection was optimal, the best F-measure in each iteration step k is plotted in Fig. 7. The plot shows that the best F-measures are not achieved in the same iterations as in phase 1 of the experiment. Table 3 compares F-measure values using iteration steps and energy thresholds from phase 1 with best F-measure values and corresponding energy thresholds and iteration steps. This presents a comparison of realistic F-measure values with upper-limit F-measure values. The last column in Table 3 shows what fraction of the upper limit was reached using a predetermined threshold and iteration step. The highest rate (89 %) was again achieved with the proposed SSA–SPA diffusion model using clique-based initial values. However, the highest F-measure value could have been reached using the proper iteration step and threshold with the basic SPA diffusion model using proposed clique-based initial values. Compared with phase 1, where users from a connected component of the mobile network graph were analysed, here we analysed very different users distributed in various parts of the network. These users could be further segmented into homogeneous subgroups of users and evaluated separately. Taking this approach, we could determine segments of users where a specific diffusion model with a specific parameter set performs best.

One might argue why the lowest number of executed iterations in Figs. 6 and 7 is equal to zero. These points present best F-measures when initial values themselves were used as churn prediction scores, before even running the iterative diffusion algorithm. The F-measure value of clique-based initial energy values used as prediction scores is already comparable to the highest best F-measure values of both diffusion models when using basic initial values. Using clique-based initial energy values as prediction scores is therefore a considerably simpler yet still relatively effective approach to churn prediction. However, to achieve better performance, it is still necessary to execute the diffusion algorithm.

Tables 2 and 3 reveal relatively small F-measure values, which are a consequence of a high number of false positives (i.e., non-churners predicted as churners) and consequently low precision. The main reason for this is the fact that there are numerous different reasons for churn and diffusion models only explain a smaller ratio of churns that are initiated by social influence. To better understand the true value of our model compared with other models, we present a lift curve plot of all four diffusion models,

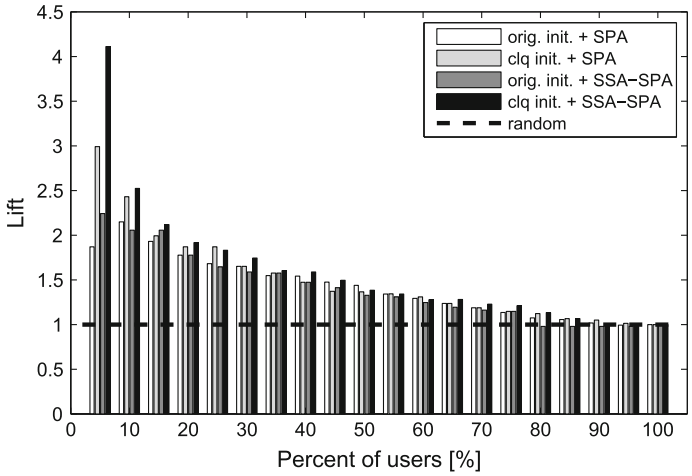

Fig. 8 Lift values for four diffusion models in the 5th, 10th, 15th, …, 100th percentiles

using iteration steps from phase 1, in Fig. 8. Lift is also a common evaluation metric when evaluating churn prediction models. The bars in Fig. 8 show the factor (lift) by which each of the four models outperforms a random assignment of churners where lift is equal to 1. A good classifier provides a high lift when only a small percentage of users are selected. Therefore, we compare our models only at the fifth percentile in Fig. 8. The best result in this percentile is yielded by our proposed SSA–SPA diffusion model using clique-based initial values, achieving a lift of 4.1. This result means that if the company would decide to offer incentives to as many future churners as possible, but it could only afford to contact 5 % of users, then using our model, they would successfully contact $5\,\% \times 4.1 = 20.5\,\%$ of all churners in a dataset. In comparison, using a model with clique-based initial values and basic SPA, they would find $5\,\% \times 3.0 = 15\,\%$ of churners. A model with basic initial values and SSA–SPA would find $5\,\% \times 2.2 = 11\,\%$ of churners, and basic SPA with basic initial values would yield $5\,\% \times 1.9 = 9.5\,\%$ of churners. The results clearly show that all three modified diffusion models outperform the basic SPA diffusion model, with the best model (i.e., SSA–SPA with clique-based initial values) outperforming the basic SPA diffusion model by 116 % in the fifth lift percentile.

6 Conclusion and issues for future research

This work proposed a new diffusion model used for predicting customer churn in the telecommunication market. The model contributes to the literature by introducing elements of social science from psychology into an energy-spreading diffusion algorithm. Improvements of the diffusion model include modification of the initial user energy determination using sociometric clique theory and modifications of the energy distribution algorithm itself by including social status theory. Additionally, a novel diffusion model prediction scheme TUSCPS that enables running SPA for individual

users using limited computational resources was proposed. This approach also offers the possibility of parallel processing.

To evaluate the proposed modifications of the diffusion model in a real context, data obtained from the largest Slovenian mobile operator were used. Results showed significant improvement of the prediction accuracy compared with that of the basic diffusion model. Performances of all considered models were evaluated using an F-measure evaluation metric and lift curve. More specifically, using a dataset of approximately 10,000 first-degree neighbours of recent churners, our model successfully discovered 20.5 % of churners when only the top 5 % of users, based on decreasing churn prediction score, were selected. In comparison, the basic SPA diffusion model successfully discovered 9.5 % of churners by selecting the same number of users.

Still, there are possibilities of further improvement. Clique-based initial values were calculated as a sum of three different initial energy contributions without any weighting or additional balancing of these contributions. By appropriately balancing specific contributions, better results might be achieved.

Phase 2 of the experiment included evaluating the proposed diffusion models on a specific dataset of users; i.e., direct neighbours of recent churners. Possibilities arise by considering other targeted subsets of users who are considered more prone to churn; e.g., users who complain to customer service support and users who frequently make calls to users at competitive service providers. Diffusion models could also be built on subsets of users that are more homogeneous, such as users of youth or senior tariff plans, or users from specific geographic areas.

To predict churners, a simple threshold method was used to separate churners from non-churners. In our work, the thresholds for each diffusion model were determined to be the same for all users. However, the threshold could be considered to be user specific and be determined using user-specific features and an appropriate continuous dependent variable regression method.

Acknowledgments Operation part financed by the European Union, European Social Fund.

References

Alba RD (1973) A graph-theoretic definition of a sociometric clique. J Math Sociol 3(1):113–126. doi:10.1080/0022250X.1973.9989826

Baras D, Ronen A, Yom-Tov E (2012) The effect of social affinity and predictive horizon on churn prediction using diffusion modeling. Tech. Rep, IBM

Blau P (1975) Approaches to the study of social structure, 1st edn. Free Press, New York

Dasgupta K, Singh R, Viswanathan B, Chakraborty D, Mukherjea S, Nanavati AA, Joshi A (2008) Social ties and their relevance to churn in mobile telecom networks. In: Proceedings of the 11th international conference on extending database technology advances in database technology (EDBT'08). ACM Press, New York, pp 668–677. doi:10.1145/1353343.1353424

Dierkes T, Bichler M, Krishnan R (2011) Estimating the effect of word of mouth on churn and cross-buying in the mobile phone market with Markov logic networks. Decis Support Syst 51(3):361–371. doi:10.1016/j.dss.2011.01.002

Ferreira JB, Vellasco M, Pacheco MA, Barbosa CH (2004) Data mining techniques on the evaluation of wireless churn. Proceedings of the European symposium on artificial neural networks. Bruges, Belgium, pp 483–488

Gary MR, Johnson DS (1979) Computers and intractability: a guide to the theory of NP-completeness. W. H. Freeman & Co., New York

Gross JL, Yellen J (eds) (2003) Handbook of graph theory, 1st edn. CRC Press, Boca Raton

Kawale J, Pal A, Srivastava J (2009) Churn prediction in MMORPGs: a social influence based approach. In: 2009 international conference on computational science and engineering, IEEE, pp 423–428. doi:10.1109/CSE.2009.80

Kempe D, Kleinberg J, Tardos E (2005) Influential nodes in a diffusion model for social networks. In: Caires L, Italiano GF, Monteiro L, Palamidessi C, Yung M (eds) Automata, languages and programming, lecture notes in computer science, vol 3580. Springer, Berlin, pp 1127–1138. doi:10.1007/11523468_91

Luce RD, Perry AD (1949) A method of matrix analysis of group structure. Psychometrika 14(2):95–116. doi:10.1007/BF02289146

Marks PEL, Cillessen AHN, Crick NR (2012) Popularity contagion among adolescents. Soc Dev 21(3):501–521. doi:10.1111/j.1467-9507.2011.00647.x

Miguéis VL, Van den Poel D, Camanho AS, Falcãoe Cunha J (2012) Predicting partial customer churn using Markov for discrimination for modeling first purchase sequences. Adv Data Anal Classif 6(4):337–353. doi:10.1007/s11634-012-0121-3

Mokken RJ (1979) Cliques, clubs and clans. Qual Quant 13(2):161–173. doi:10.1007/BF00139635

Moody J (2001) Peer influence groups: identifying dense clusters in large networks. Soc Netw 23(4):261–283. doi:10.1016/S0378-8733(01)00042-9

Moreno JL (1953) Who shall survive? Foundations of sociometry, group psychotherapy and socio-drama, sociometry monographs, 2nd edn. Beacon House, Oxford

Mozer MC, Wolniewicz R, Grimes DB, Johnson E, Kaushansky H (2000) Predicting subscriber dissatisfaction and improving retention in the wireless telecommunications industry. IEEE Trans Neural Netw 11(3):690–696. doi:10.1109/72.846740

Nath SV, Behara RS (2003) Customer churn analysis in the wireless industry: a data mining approach. In: Proceedings-annual meeting of the decision sciences institute, pp 505–510 (2003)

Naveen N, Ravi V, Rao CR (2010) Data mining via rules extracted from GMDH: an application to predict churn in bank credit cards. In: Setchi R, Jordanov I, Howlett RJ, Jain LC (eds) Knowledge-based and intelligent information and engineering systems, vol 6276., Lecture notes in computer scienceSpringer, Berlin, pp 80–89. doi:10.1007/978-3-642-15387-7_12

Plunkett JW (2012) Plunkett's telecommunications industry almanac 2013. Plunkett Research Ltd, Houston

Powers DMW (2011) Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. J Mach Learn Technol 2(1):37–63

Qi J, Zhang L, Liu Y, Li L, Zhou Y, Shen Y, Liang L, Li H (2008) ADTreesLogit model for customer churn prediction. Ann Oper Res 168(1):247–265. doi:10.1007/s10479-008-0400-8

Richter Y, Yom-Tov E, Slonim N (2010) Predicting customer churn in mobile networks through analysis of social groups. In: Proceedings of the 2010 SIAM international conference on data mining (SDM 2010), pp 732–741. doi:10.1137/1.9781611972801.64

Rosenberg LJ, Czepiel JA (1984) A marketing approach for customer retention. J Consum Mark 1(2):45–51. doi:10.1108/eb008094

Saltelli A, Ratto M, Andres T, Campolongo F, Cariboni J, Gatelli D, Saisana M, Tarantola S (2008) Global sensitivity analysis: the primer. Wiley, Chichester. doi:10.1002/9780470725184

van den Berg YHM, Cillessen AHN (2012) Computerized sociometric and peer assessment: an empirical and practical evaluation. Int J Behav Dev 37(1):68–76. doi:10.1177/0165025412463508

Verbeke W, Dejaeger K, Martens D, Hur J, Baesens B (2012) New insights into churn prediction in the telecommunication sector: a profit driven data mining approach. Eur J Oper Res 218(1):211–229. doi:10.1016/j.ejor.2011.09.031

Yan L, Wolniewicz RH, Dodier R (2004) Predicting customer behavior in telecommunications. Intell Syst IEEE 19(2):50–58

Yoon S, Koehler J, Ghobarah A (2010) Prediction of advertiser churn for google adwords. In: JSM proceedings (2010)

Zhang X, Zhu J, Xu S, Wan Y (2012) Predicting customer churn through interpersonal influence. Knowl Based Syst 28:97–104. doi:10.1016/j.knosys.2011.12.005

Zhang X, Liu Z, Yang X, Shi W, Wang Q (2010) Predicting customer churn by integrating the effect of the customer contact network. In: Proceedings of 2010 IEEE international conference on service operations and logistics, and informatics. IEEE, pp 392–397 (2010). doi:10.1109/SOLI.2010.5551545